Comprehensive evaluation of gene sequence encoding methods in deep learning

Han LI[1], Jiming HU[1]& Xiaoyong SUN[1,*]

[1]Agricultural Big-Data Research Center, College of Information Science and Engineering, Shandong Agricultural University, Tai'an 271018, China.

[*]Corresponding author

Han LI：collected and analyzed the data, performed the experiment, wrote and revised the manuscript;
Jiming HU：performed the experiment;
Xiaoyong SUN：proposed the research problems, designed research program.

Abstract

Background: The prediction of genomic structure has become a hot spot in genome research. At present, the prediction method based on deep learning is more effective and accurate than other machine learning algorithms. Since gene sequence data cannot directly enter the deep learning model, the original data need to be encoded and converted into numerical features before model prediction. As a result, different encoding methods may affect final accuracy.

Methods: In order to explore the performance of different encoding methods, we compared ten strategies in six deep learning models. We also compared the performance of all methods on independent datasets and models from our laboratory. For all models, we used their original parameters.

Results: Dummy encoding, hash encoding, and one-hot encoding perform best in various models. In addition, dummy encoding and one-hot encoding are the best for processing RNA data, while hash encoding is superior to other methods for processing promoter data. Also, when processing part- or full-sequence data, the performance of dummy encoding, hash encoding, and one-hot encoding is similar. Besides that, in sisRNA datasets and prediction models of *Arabidopsis* and rice, dummy encoding and one-hot encoding achieve higher prediction accuracy.

Conclusions: We conclude that the best encoding method varies when the data set changes. One-hot encoding, dummy encoding, and hash encoding are the three best methods for six models. This study fills the gap on sequence encoding methods in deep learning and can provide a valuable reference for the community.

Keywords: deep learning; RNA; promoter; encoding methods

1. Introduction

Genomics is the “microscope” for human beings to explore the mysteries of life. Various genomic features help the community to understand the mechanisms behind life. The development of high-throughput sequencing technology has generated a large amount of gene data, which encourages people to study genomic structure [1]. Existing research shows that some important genomic structures, such as long non-coding RNAs (lncRNAs), circular RNAs (circRNAs), extrachromosomal circular DNA (eccDNA), and promoters, play a key role in regulating biological and life activities[2-6]. Therefore, recognition and prediction of important genomic structures can address major problems in biology, medicine, and agriculture[7].

At present, machine learning and deep learning have been widely used in the prediction of genomic structure. As long as biological sequence data are converted into numerical features, a model can automatically predict the structure[8-10]. Currently, many algorithms, including the Markov model[11], second-order Markov model[12], hidden Markov model[13,14], pseudo dinucleotide composition (PseDNC) [15-17], k-spectrum nucleotide pair frequency (KSNPF)[18], and k-space nucleotide composition (KSNC)[19,20], have been widely used in prediction of genomic structures based on machine learning, but the limitation of machine learning is that it needs to extract features manually. Deep learning solves this problem. The advantage of deep learning is that a model can automatically learn features, thereby achieving the prediction of genomic structures[21-23]. At present, one-hot encoding[24] is prevalent in almost all researches, and many mature sequence encoding methods based on deep learning have been produced, such as k-mer[25], word vector[26], etc. Despite many encoding methods having been applied to deep learning models, most of the existing researches usually use a certain method directly. Different encoding methods may cause different features to be learned by models, which leads to differences in model prediction accuracy. Hence, it is necessary to explore the effects of different sequence encoding methods on the prediction results of deep learning models.

In this study, we compared ten different encoding methods: baseN encoding, binary encoding, complementary encoding, dummy encoding, effect encoding, hash encoding, index encoding, one-bit encoding, one-hot encoding, and ternary encoding. We evaluated the performance of the ten methods based on the prediction accuracy of models. In addition, we also conducted experiments on independent datasets and models. Compared with the other methods, dummy encoding, hash encoding, and one-hot encoding perform best. Dummy encoding and one-hot encoding are more suitable for encoding RNA data, while hash encoding is more suitable for encoding promoter data, and the three encoding methods perform equally on part- or full-sequence data. What's more, dummy encoding and one-hot encoding also perform best on stable intronic sequence RNA (sisRNA)[27] datasets and prediction models of *Arabidopsis* and rice. The above results show that in addition to one-hot encoding, dummy encoding and hash encoding can also enable the model to learn more features, so as to achieve higher prediction accuracy (Fig. 1).

2. Methodology

2.1. Datasets

The five sequence datasets used in this study were derived from existing studies; the data are described in detail below (Table 1):

1) circRNAs[28], downloaded from circBase (http://www.circbase.org/). After preprocessing, the final sequence length is 80 bp.

2) Linear RNAs, downloaded from EMBL-EBI (https://www.ebi.ac.uk/), including *Arabidopsis thaliana*, maize, rice, and their mixed data; after preprocessing, the final sequence length is 200 bp.

3) *Zea mays* lncRNAs are from Meng et al. (2021)[29]; the data use the longest sequence length as the standard and perform 0-padding on the sequence whose length is less than it.

4) The *E. coli* promoter is from Shujaat et al. (2020)[30], and the sequence length is 81 bp.

5) The *Saccharomyces cerevisiae* promoter is from Vaishnav et al. (2022)[31]; the sequence length is 110 bp. Owing to the limitation of hardware conditions, 100,000 sequences were randomly selected from this study.

2.2. Encoding methods

A total of ten sequence encoding methods are used in this paper; all encoding methods and experimental procedures are shown in Fig. 2.

2.2.1. One-bit encoding

This method was proposed in 2012 by Church et al.[32], and only one bit is encoded for each base in the DNA sequence, so we call it one-bit encoding. With one-bit encoding, A is encoded as 0, T is encoded as 1, G is encoded as 1, and C is encoded as 0.

2.2.2. Index encoding

Index encoding converts discrete type features into continuous numerical variables and can normalize discrete data so that each discrete category has a separate index. Using index encoding, A can be encoded as 1, T as 2, G as 3, and T as 4.

2.2.3. Complementary encoding

The bases in a DNA sequence are paired complementary. Therefore, to imitate this principle, we propose complementary encoding to try to encode complementary bases. We encode the complementary bases with the opposite numbers. With complementary encoding, A can be encoded as 1, T as −1, G as 2, and C as −2.

2.2.4. BaseN encoding

BaseN encoding represents base N, using N unique numbers to represent all elements in the sequence, reducing the number of features that efficiently represent data, and improving memory usage. Using baseN encoding, A is encoded as (0, 1), T is encoded as (0, 2), G is encoded as (0, 3), and C is encoded as (0, 4).

2.2.5. Dummy encoding

Dummy encoding is a method of assigning classified variables, which converts DNA sequence data into a set of binary variables. Different from one-hot encoding, dummy encoding selects one category as a reference for variables with n classification attributes to generate $n-1$ categories, that is, it uses $n-1$ features to represent n categories of data. Using dummy encoding, A can be encoded as (0, 0, 1), T as (0, 1, 0), G as (1, 0, 0), and C as (0, 0, 0).

2.2.6. Ternary encoding

Ternary encoding is similar to binary encoding, but ternary logic is closer to the way the human brain thinks. Ternary encoding represents data by 0, 1, 2. Using ternary encoding, A can be encoded as (0, 0, 1), T as (0, 0, 2), G as (0, 1, 0), and C as (0, 1, 1).

2.2.7. Effect encoding

Effect encoding is similar to dummy encoding, using $n - 1$ features to represent n categories of data, but different from dummy encoding, lines that contain only 0 in dummy encoding are replaced by -1 in effect encoding. Using effect encoding, A can be encoded as (0, 0, 1), T as (0, 1, 0), G as (1, 0, 0), and C as (-1, -1, -1).

2.2.8. Binary encoding

In binary encoding, categorical features are converted into different numerical values by an ordinal encoder, and these numerical values are then converted into binary numbers. Using binary encoding, A can be encoded as (0, 0, 0, 1), T as (0, 0, 1, 0), G as (0, 0, 1, 1), and C as (0, 1, 0, 0).

2.2.9. One-hot encoding

One-hot encoding, also known as one-bit efficient encoding, uses n features to represent n categories of data. One-hot encoding is one of the most commonly used sequence encoding methods in the field of bioinformatics. Using one-hot encoding, A can be encoded as (0, 0, 0, 1), T can be encoded as (0, 0, 1, 0), G can be encoded as (0, 1, 0, 0), and C can be encoded as (1, 0, 0, 0).

2.2.10. Hash encoding

Hash is an algorithm that converts an arbitrary-length input to a fixed-length output; like one-hot encoding, it uses new dimensions to represent categorical features. But in hash encoding, no matter how many categories the data have, they can be represented by n new features. Using hash encoding, A can be encoded as (0, 0, 0, 0, 1), T as (0, 0, 0, 1, 0), G as (0, 0, 1, 0, 0), and C as (0, 1, 0, 0, 0).

2.3. Models

In this study, we used six models, two of which were previously proposed in our laboratory, namely DeepCircRNA (http://deepbiology.cn/DeepCircRNA) and DeepAS (http://deepbiology.cn/DeepAS). DeepCircRNA is a model based on a Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM). DeepAS is a model based on CNN, LSTM, and Gated Recurrent Unit (GRU). This

study also compared the ten encoding methods in four published models from other laboratories: SpliceFinder[33], PlncRNA-HDeep[29], pcPromoter-CNN[30], and Evolution-gpu [31].

2.4. Performance evaluation

In order to measure the performance of the different encoding methods in preserving sequence information, we used accuracy as an indicator to evaluate the advantages and disadvantages of each encoding method. In addition, we also used the original evaluation indicators of each model to evaluate them, including precision, recall, F1-score, etc. The calculation formulas for all indicators are:

$$accuracy = (TP+TN)/(TP+FP+TN+FN) \quad (1)$$

$$precision = TP/(TP+FP) \quad (2)$$

$$recall = TP/(TP+FN) \quad (3)$$

$$F1\text{-}score = 2TP/(2TP+FP+FN) \quad (4)$$

where TP represents the number of true positive samples, TN represents the number of true negative samples, FP represents the number of false positive samples, and FN represents the number of false negative samples.

3. Results

In the following studies, we only discuss the influence of different encoding methods on the prediction accuracy of models, thus we use all the original parameters of the models.

3.1. Performance of ten encoding methods in six models

3.1.1. Model training

We used ten encoding methods to process all sequence data and input them into the corresponding model for training. For DeepCircRNA, when using baseN encoding, index encoding, and one-bit encoding, the initial accuracy of the model was the lowest, while the initial accuracy of the model using both hash encoding and one-hot encoding was higher than 0.7. On increasing the number of epochs, the accuracy of the model began to grow slowly, and finally stabilized. After the training, the accuracy of baseN encoding, index encoding, and one-bit encoding was obviously lower than that of the other methods, while one-hot encoding made the model achieve the highest accuracy (Fig. 3A).

For DeepAS, in all datasets, the initial accuracy of baseN encoding, index encoding, and one-bit encoding was the lowest. Dummy encoding, hash encoding, and one-hot encoding could achieve the highest accuracy in the fewest epochs, while the accuracy of baseN encoding, index encoding, and one-bit encoding was obviously lower than that of the other methods, and the performance of baseN encoding and index encoding was very unstable (Fig. 3B). For SpliceFinder, since we cannot obtain the SpliceFinder datasets, and the work of SpliceFinder is similar to our previous study in DeepAS, we used all the datasets of DeepAS to train and test SpliceFinder. In all datasets, the initial accuracy of baseN encoding, index encoding, and one-bit encoding was the lowest, and after the training, the accuracy of these three methods was also significantly lower than that of the other methods (Fig. 3C). For PlncRNA-HDeep, the accuracy of one-bit encoding in the whole training process was always lower than for the other methods (Fig. 3D). For Evolution-gpu, the Pearson Correlation Coefficient (PCC) of one-bit encoding in the whole training process was obviously lower than that for the other methods (Fig. 3E). Therefore, in the training process of the above models, baseN encoding, index encoding, and one-bit encoding were the most unstable, and the accuracy was the lowest after the model training.

3.1.2. Model testing

For DeepCircRNA, one-hot encoding had the highest accuracy; the accuracy of index encoding was the lowest (Table 2). For DeepAS, dummy encoding had the highest accuracy in *Arabidopsis* and mixed datasets, and the accuracy of one-hot encoding was the highest in maize and rice datasets (Table 3). It is worth noting that the performance of baseN encoding and index encoding was the most unstable. In addition, the accuracy of one-bit encoding was also lower than that of the other methods. For SpliceFinder, dummy encoding achieved the highest accuracy in *Arabidopsis*, rice, and mixed datasets, while one-hot encoding had the highest accuracy in maize datasets (Supplementary Table 1). For PlncRNA-HDeep, one-hot encoding achieved the highest accuracy, F1-score, and recall, while dummy encoding achieved the highest precision (Supplementary Table 2). For pcPromoter-CNN, after 5-fold cross validation, hash encoding achieved the highest accuracy, and the Area Under the ROC Curve (AUC) was also the highest of all methods (Supplementary Table 3). For Evolution-gpu, the

PCC of hash encoding was the highest (Supplementary Table 4). Based on all the above results, we draw the conclusion that dummy encoding, hash encoding, and one-hot encoding perform best. With these three encoding methods, models can learn more features to achieve better prediction results.

3.2. Performance of ten encoding methods with different types of data

There are two kinds of datasets used in our work, RNA datasets and promoter datasets. For RNA datasets, one-hot encoding performed best in human, maize, rice (for DeepAS), and *Zea mays* datasets, and dummy encoding performed best in *Arabidopsis*, rice (for SpliceFinder), and mixed datasets (Fig. 4A). For promoter datasets, hash encoding achieved the best prediction effect (Fig. 4B). Therefore, we consider that dummy encoding and one-hot encoding are more suitable for processing RNA data, and hash encoding is more suitable for processing promoter data.

3.3. Comparison of ten encoding methods with part-sequence and full-sequence data

Since all sequences of original data are not equal in length, the original sequence needs to be preprocessed before encoding. In the datasets used in this study, in order to ensure consistency of data length, all data were preprocessed, that is, by sequence intercepting or sequence padding. We regard the preprocessed data as a part sequence or full sequence. A part sequence is a sequence of a certain length intercepted from the whole sequence according to a certain rule, while a full sequence is a sequence of a certain length as a benchmark; for a sequence less than this length, 0-padding is performed until it reaches this length. Therefore, in the data used in this study, human, *Arabidopsis*, maize, rice, mixed, and *E. coli* are part-sequence data, *Zea mays* is full-sequence data, and the *Saccharomyces cerevisiae* promoter has both part- and full-sequence data. For part-sequence data, dummy encoding, hash encoding, and one-hot encoding performed best (Supplementary Fig. 1A). For full-sequence data, one-hot encoding performed best, and hash encoding and dummy encoding followed closely (Supplementary Fig. 1B). For data with both part and full sequences, hash encoding performed best, followed by one-hot encoding and dummy encoding (Supplementary Fig. 1C). Therefore, we conclude that for part- and full-sequence data, the performance of the above three encoding methods is similar.

3.4. Testing of ten encoding methods on independent datasets and model

In order to verify the performance of the ten encoding methods on other genomic structures, we conducted experiments on independent datasets and models. We used sisRNA datasets detected in our laboratory, and input them into the newly constructed sisRNA prediction model in our laboratory. For *Arabidopsis* and rice, the training set sample is 5000, and the test set sample is 1000; 20% is divided from the training set as the validation set, and the rest is the final training set. The ratio of positive and negative samples is 1:1; the *Arabidopsis* sequence length is 200 bp and the rice sequence length is 600 bp. The model used is CNN combined with GRU. For *Arabidopsis*, dummy encoding had the highest prediction accuracy, while one-hot encoding had the highest prediction accuracy in rice (Supplementary Table 5).

4. Discussion

One-hot encoding is one of the most widely used encoding methods in previous studies. In this study, one-hot encoding also showed excellent performance. This may be due to one-hot encoding using n binary characteristics to represent n categories, from which the model can learn regular patterns. However, one-hot encoding is not an irreplaceable method, and other methods may achieve a similar effect. Therefore, it is necessary to explore other simple and effective encoding methods, which can provide more possibilities for sequence encoding.

In our study, dummy encoding and one-hot encoding perform best in RNA data, while hash encoding performs best in promoter data. What's more, dummy encoding, hash encoding, and one-hot encoding perform very well in both part-sequence data and full-sequence data. All these indicate that one-hot encoding is not the only encoding method suitable for sequence data. Instead, dummy encoding and hash encoding may help a model learn more features to achieve excellent prediction accuracy.

We used ten encoding methods and six models to evaluate the model performance. However, there are some limitations of this research. Firstly, we only conducted experiments on RNA and promoters. In the future, we will test our finding on more genomic features (including enhancers, tandem repeats, transposons, DNA methylation, proteins, etc.). Secondly, in this study, we chose five species for model prediction. In the next step, we plan to include more species to enrich our knowledge of these encoding algorithms.

5. Conclusion

In this paper, we compare the performance of ten encoding methods in six models. We evaluated the performance of all methods based on prediction accuracy. We also compared ten methods on independent datasets and model. The results indicate that dummy encoding, hash encoding, and one-hot encoding can make the model achieve higher prediction accuracy. In particular, for RNA, the

accuracy of dummy encoding and one-hot encoding is the highest, and for promoters, hash encoding performs best. Furthermore, in *Arabidopsis* and rice sisRNA datasets, dummy encoding and one-hot encoding, respectively, make the model achieve the highest prediction accuracy. Our study has filled the gap in the research on sequence encoding methods and provided a reference for experts in the field of bioinformatics to predict the important genomic structures.

ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (grant number 32070684 to X.S.). We thank Supercomputing Center in Shandong Agricultural University for technical support.

References

[1] K.D. Christensen, D. Dukhovny, U. Siebert, R.C Green, Assessing the costs and Cost-Effectiveness of genomic sequencing, J. Pers. Med. 5 (4) (2015) 470-486.

[2] B.L. Gudenas, L. Wang, Prediction of lncRNA subcellular localization with deep learning from sequence features, Sci Rep 8 (1) (2018) 16385.

[3] S. Wang, X. Cheng, Y. Li, M. Wu, Y. Zhao, Image-based promoter prediction: a promoter prediction method based on evolutionarily generated patterns, Sci Rep 8 (1) (2018) 17695.

[4] J. Ye, L. Wang, S. Li, Q. Zhang, Q. Zhang, W. Tang, K. Wang, K. Song, G. Sablok, X. Sun, H. Zhao, AtCircDB: a tissue-specific database for Arabidopsis circular RNAs, Brief Bioinform 20 (1) (2019) 58-65.

[5] K. Wang, H. Tian, L. Wang, L. Wang, Y. Tan, Z. Zhang, K. Sun, M. Yin, Q. Wei, B. Guo, J. Han, P. Zhang, H. Li, Y. Liu, H. Zhao, X. Sun, Deciphering extrachromosomal circular DNA in Arabidopsis, Comput Struct Biotechnol J. 19 (2021) 1176-1183.

[6] A.M. Oudelaar, D.R. Higgs, The relationship between genome structure and function, Nat Rev Genet 22 (3) (2021) 154-168.

[7] W.N. Moss, J.A. Steitz, Genome-wide analyses of Epstein-Barr virus reveal conserved RNA structures and a novel stable intronic sequence RNA, BMC Genom 14 (2013) 543.

[8] W.Y. He, C.Z. Jia, Q. Zou, 4mCPred: machine learning methods for DNA N4-methylcytosine sites prediction, Bioinformatics 35 (4) (2019) 593-601.

[9] M.P. Niu, J. Wu, Q. Zou, Z.D. Liu, L. Xu, rBPDL:Predicting RNA-binding proteins using deep learning, IEEE J Biomed Health Inform 25 (9) (2021) 3668-3676.

[10] Y.P. Lei, S.Y. Li, Z.Y. Liu, F.P. Wan, T.Z. Tian, S. Li, D. Zhao, J.Y. Zeng, A deep-learning framework for multi-level peptide-protein interaction prediction, Nat Commun 12 (1) (2021) 5465.

[11] C. Pian, G. Zhang, F. Li, X.D. Fan, MM-6mAPred: identifying DNA N6-methyladenine sites based on Markov model, Bioinformatics 36 (2) (2020) 388-392.

[12] J. Yang, K. Lang, G. Zhang, X. Fan, Y. Chen, C. Pian, SOMM4mC: a second-order Markov model for DNA N4-methylcytosine site prediction in six species, Bioinformatics 36 (14) (2021) 4103-4105.

[13] T.M. Ji, A Bayesian hidden Markov model for detecting differentially methylated regions, Biometrics 75 (2) (2019) 663-673.

[14] A. Dhar, D.K. Ralph, V.N. Minin, F.A. Matsen, A Bayesian phylogenetic hidden Markov model for B cell receptor sequence analysis, PLoS Comput Biol 16 (8) (2020) e1008030.

[15] W. Chen, H. Ding, X. Zhou, H. Lin, K.C. Chou, iRNA(m6A)-PseDNC: Identifying N6-methyladenosine sites using pseudo dinucleotide composition, Anal Biochem 561-562 (2018) 59-65.

[16] J. Song, J.J. Zhai, E. Bian, Y.J. Song, J.T. Yu, C. Ma, Transcriptome-Wide annotation of m5C RNA modifications using machine learning, Front Plant Sci 9 (2018) 519.

[17] T. Fang, Z.Z. Zhang, R. Sun, L. Zhu, J.J. He, B. Huang, Y. Xiong, X.L. Zhu, RNAm5CPred: Prediction of RNA 5-methylcytosine sites based on three different kinds of nucleotide composition, Mol Ther Nucleic Acids 18 (2019) 739-747.

[18] X. Chen, Y. Xiong, Y.B. Liu, Y.Q. Chen, S.D. Bi, X.L. Zhu, m5CPred-SVM: a novel method for predicting m5C sites of RNA, BMC Bioinform 21 (1) (2020) 489.

[19] M.M. Hasan, B. Manavalan, W. Shoombuatong, M.S. Khatun, H. Kurata, i4mC-Mouse: Improved identification of DNA N4-methylcytosine sites in the mouse genome using multiple encoding schemes, Comput Struct Biotechnol J 18 (2020) 906-912.

[20] M.M. Hasan, B. Manavalan, W. Shoombuatong, M.S. Khatun, H. Kurata, i6mA-Fuse: improved and robust prediction of DNA 6 mA sites in the Rosaceae genome by fusing multiple feature representation, Plant Mol Biol 103 (1-2) (2020) 225-234.

[21] Z. Abbas, H. Tayara, K.T. Chong, 4mCPred-CNN-Prediction of DNA N4-methylcytosine in the mouse genome using a convolutional neural network, Genes (Basel). 12 (2) (2021) 296.

[22] T.H. Yang, S.C. Shiue, K.Y. Chen, Y.Y. Tseng, W.S. Wu, Identifying piRNA targets on mRNAs in C. elegans using a deep multi-head attention network, BMC Bioinform 22(1) (2021) 503.

[23] Y. Li, F.K. Kong, H. Cui, F. Wang, C.Q. Li, J.Q. Ma, SENIES: DNA shape enhanced two-layer deep learning predictor for the identification of enhancers and their strength, IEEE/ACM Trans Comput Biol Bioinform (2022) PP.

[24] X.M. Zheng, S.G. Xu, Y. Zhang, X.X. Huang, Nucleotide-level convolutional neural networks for pre-miRNA classification, Sci Rep 9 (1) (2019) 628.

[25] M. Tahir, M. Hayat, K.T. Chong, Prediction of N6-methyladenosine sites using convolution neural network model based on distributed feature representations, Neural Netw 129 (2020) 385-391.

[26] J.W. Hong, R.T. Gao, Y. Yang, CrepHAN: Cross-species prediction of enhancers by using hierarchical attention networks, Bioinformatics 37 (20) (2021) 3436-3443.

[27] S.N. Chan, J.W. Pek, Stable intronic sequence RNAs (sisRNAs): an expanding universe, Trends Biochem Sci. 44 (3) (2018) 258-272.

[28] Glažar P, Papavasileiou P, Rajewsky N. circBase: a database for circular RNAs. RNA. 2014 . 20(11):1666-1670.

[29] J. Meng, Q. Kang, Z. Chang, Y.S. Luan, PlncRNA-HDeep: plant long noncoding RNA prediction using hybrid deep learning based on two encoding styles, BMC Bioinform 22 (Suppl 3) (2021) 242.

[30] M. Shujaat, A. Wahab, H. Tayara, K.T. Chong, pcPromoter-CNN: a CNN-based prediction and classification of promoters, Genes (Basel) 11 (12) (2020) 1529

[31] E.D. Vaishnav, C.G. de Boer, J. Molinet, M. Yassour, L. Fan, X. Adiconis, D.A. Thompson, J.Z. Levin, F.A. Cubillos, A. Regev, The evolution, evolvability and engineering of gene regulatory DNA, Nature 603 (7901) (2022) 455-463.

[32] G.M. Church, Y. Gao, S. Kosuri, Next-generation digital information storage in DNA, Science 337 (6102) (2012) 1628.

[33] R.H. Wang, Z.S. Wang, J.P. Wang, S.C. Li, SpliceFinder: ab initio prediction of splice sites using convolutional neural network, BMC Bioinform 20 (Suppl 23) (2019) 652.

Figure Legends

Fig. 1 The best encoding methods of all data. We assign a value of 1 to the encoding method with the best performance, and 0 to other methods. The heatmap indicates that hash encoding, dummy encoding and one-hot encoding are the best methods.

Fig. 2 Ten encoding methods and complete experimental process. M1: one bit encoding; M2: index encoding; M3: complementary encoding; M4: baseN encoding; M5: dummy encoding; M6: ternary encoding; M7: index encoding; M8: binary encoding; M9: one-hot encoding; M10: hash encoding.

Fig. 3 Training process of models. In figure legend, At the top is the method to maximize the prediction accuracy (or PCC) of the model. A. Training process of human circRNAs using ten encoding methods in DeepCircRNA. B. The training process of *Arabidopsis*, maize, rice and their mixed datasets using ten encoding methods in DeepAS. C. The training process of *Arabidopsis*, maize, rice and their mixed datasets using ten encoding methods in SpliceFinder. D. Training process of *Zea mays* lncRNAs using ten encoding methods in PlncRNA-HDeep. E. Training process of *Saccharomyces cerevisiae* promoter using ten encoding methods in Evolution-gpu.

Fig. 4 Performance of ten enoding methods in different types of data. In figure legend, the method at the top has the highest prediction accuracy (or PCC). A. Performance of ten encoding methods in RNA datasets. Radial bar charts show that the prediction accuracy of dummy encoding and one-hot encoding is the highest. B. Performance of ten encoding methods in promoter datasets. Radial bar charts indicate that hash encoding can achieve the highest accuracy (or PCC).

Supplementary Fig. 1 Performance of ten encoding methods in full sequence data and part sequence data. In this figure, the method with the highest prediction accuracy (or PCC) is at the top. A. Performance of all encoding methods in part sequence data. Funnel plots show that the best performance is dummy encoding, hash encoding and one-hot encoding. B. Performance of all encoding methods in full sequence data. Funnel plots show that one-hot encoding performs best, and hash encoding, dummy encoding follow closely. C. Performance of all encoding methods in data with

both part and full sequence. Funnel plots show that hash encoding performs best, followed by one-hot encoding and dummy encoding.

Figures

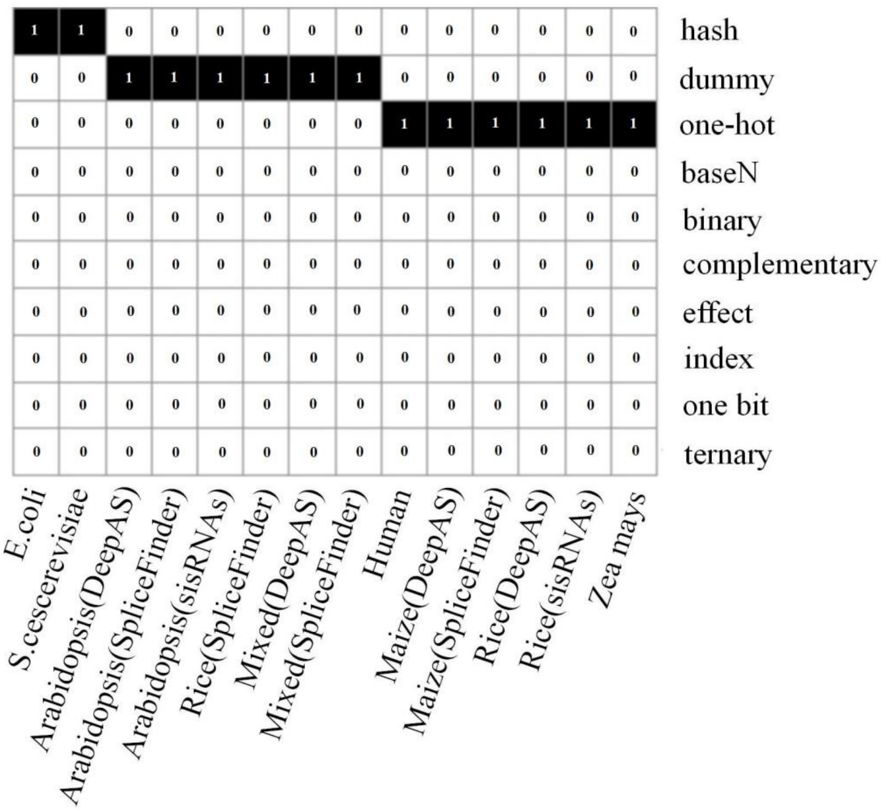


Fig. 1. The best encoding methods of all data. We assign a value of 1 to the encoding method with the best performance, and 0 to other methods. The heatmap indicates that hash encoding, dummy encoding and one-hot encoding are the best methods.

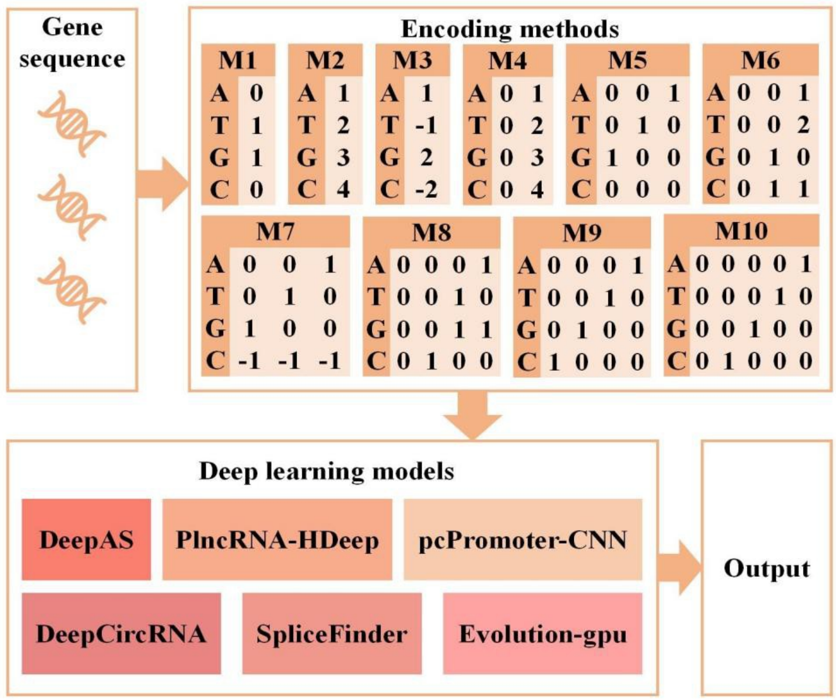
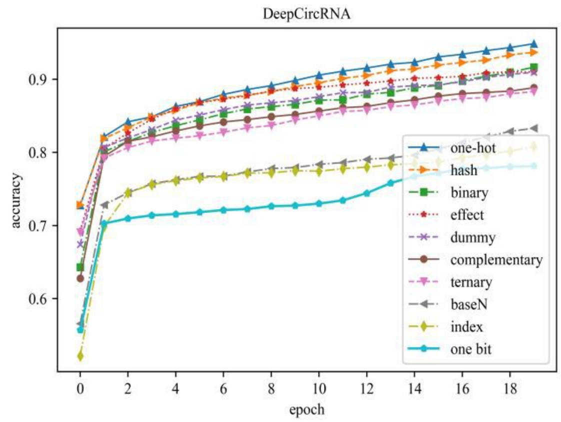
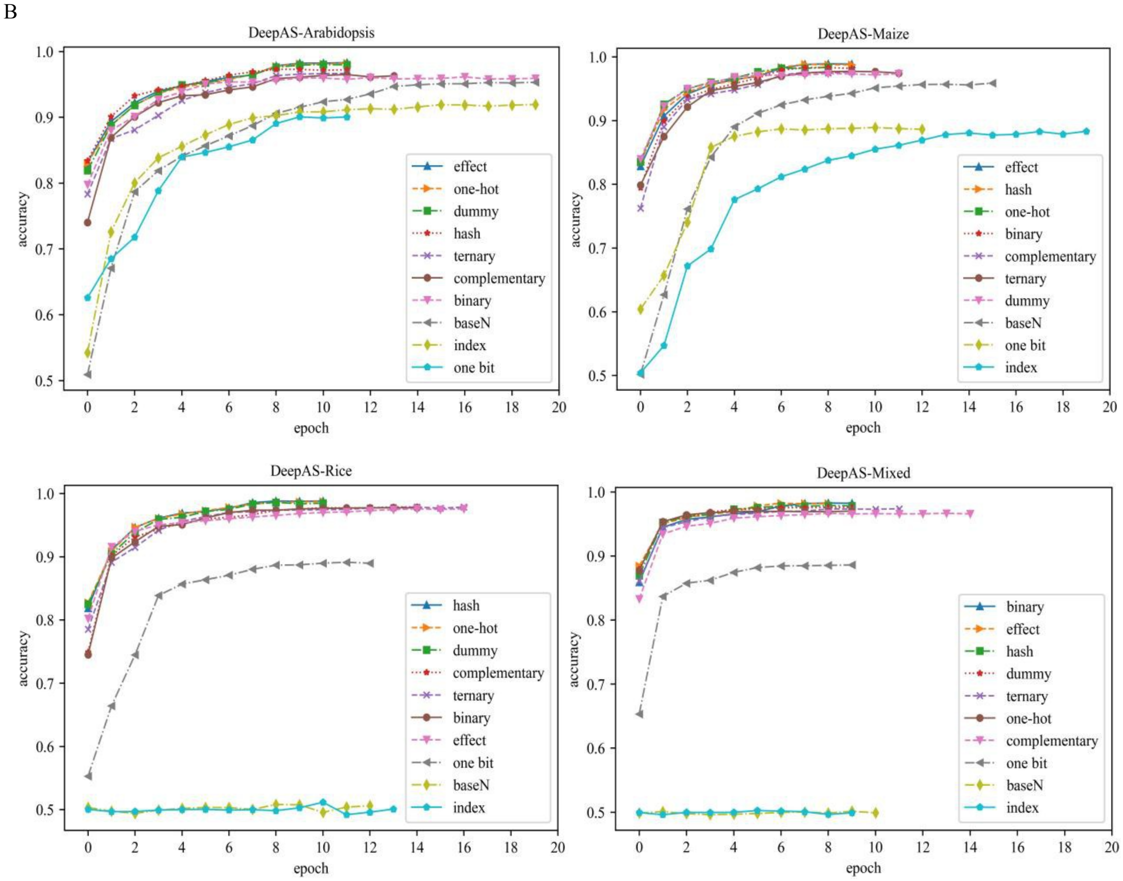


Fig. 2. Ten encoding methods and complete experimental process. M1: one bit encoding; M2: index encoding; M3: complementary encoding; M4: baseN encoding; M5: dummy encoding; M6: ternary encoding; M7: index encoding; M8: binary encoding; M9: one-hot encoding; M10: hash encoding.

A


DeepCircRNA
accuracy
epoch
one-hot
hash
binary
effect
dummy
complementary
ternary
baseN
index
one bit

B


DeepAS-Arabidopsis
accuracy
epoch
effect
one-hot
dummy
hash
ternary
complementary
binary
baseN
index
one bit
DeepAS-Maize
accuracy
epoch
effect
hash
one-hot
binary
complementary
ternary
dummy
baseN
one bit
index
DeepAS-Rice
accuracy
epoch
hash
one-hot
dummy
complementary
ternary
binary
effect
one bit
baseN
index
DeepAS-Mixed
accuracy
epoch
binary
effect
hash
dummy
ternary
one-hot
complementary
one bit
baseN
index

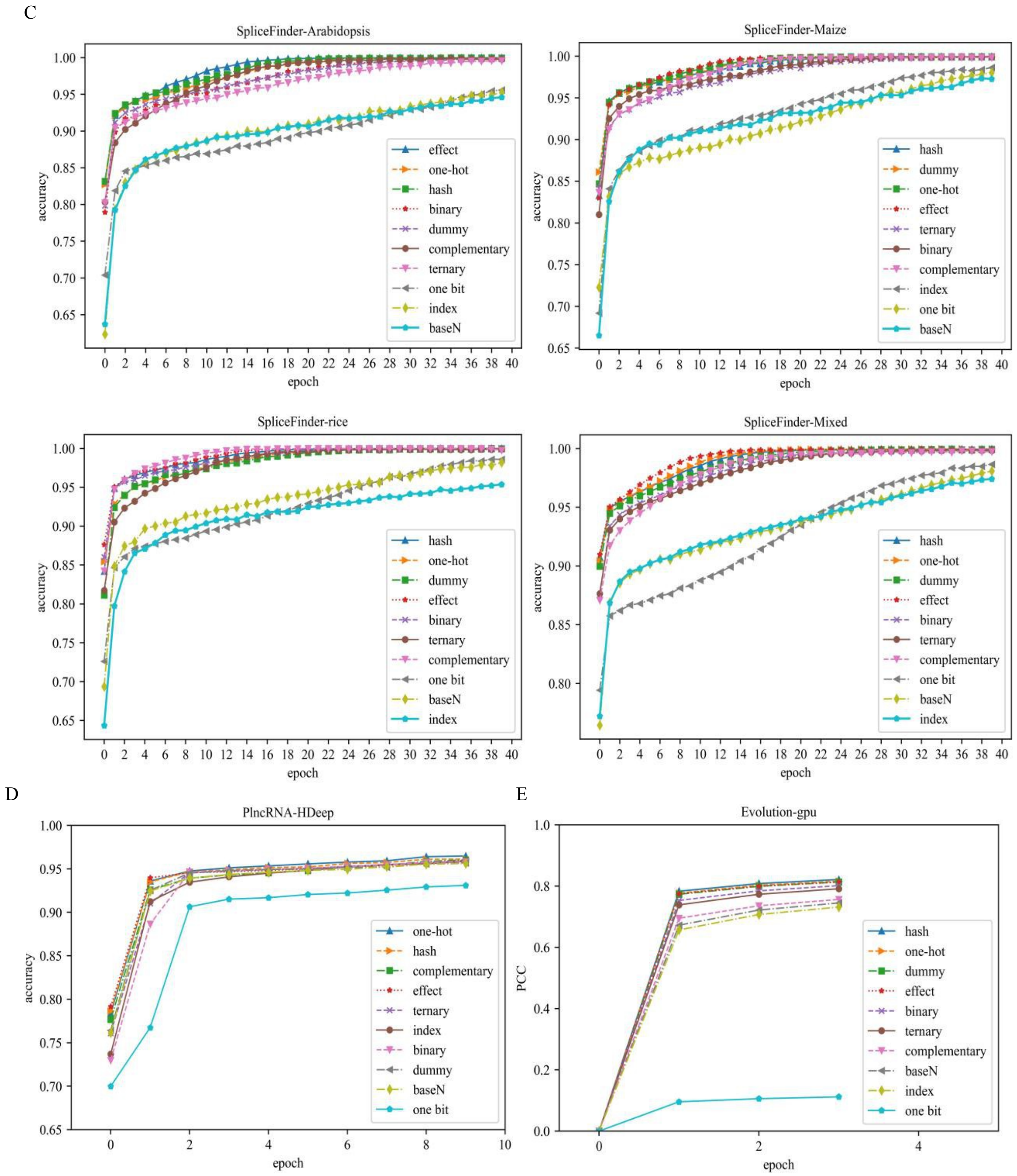
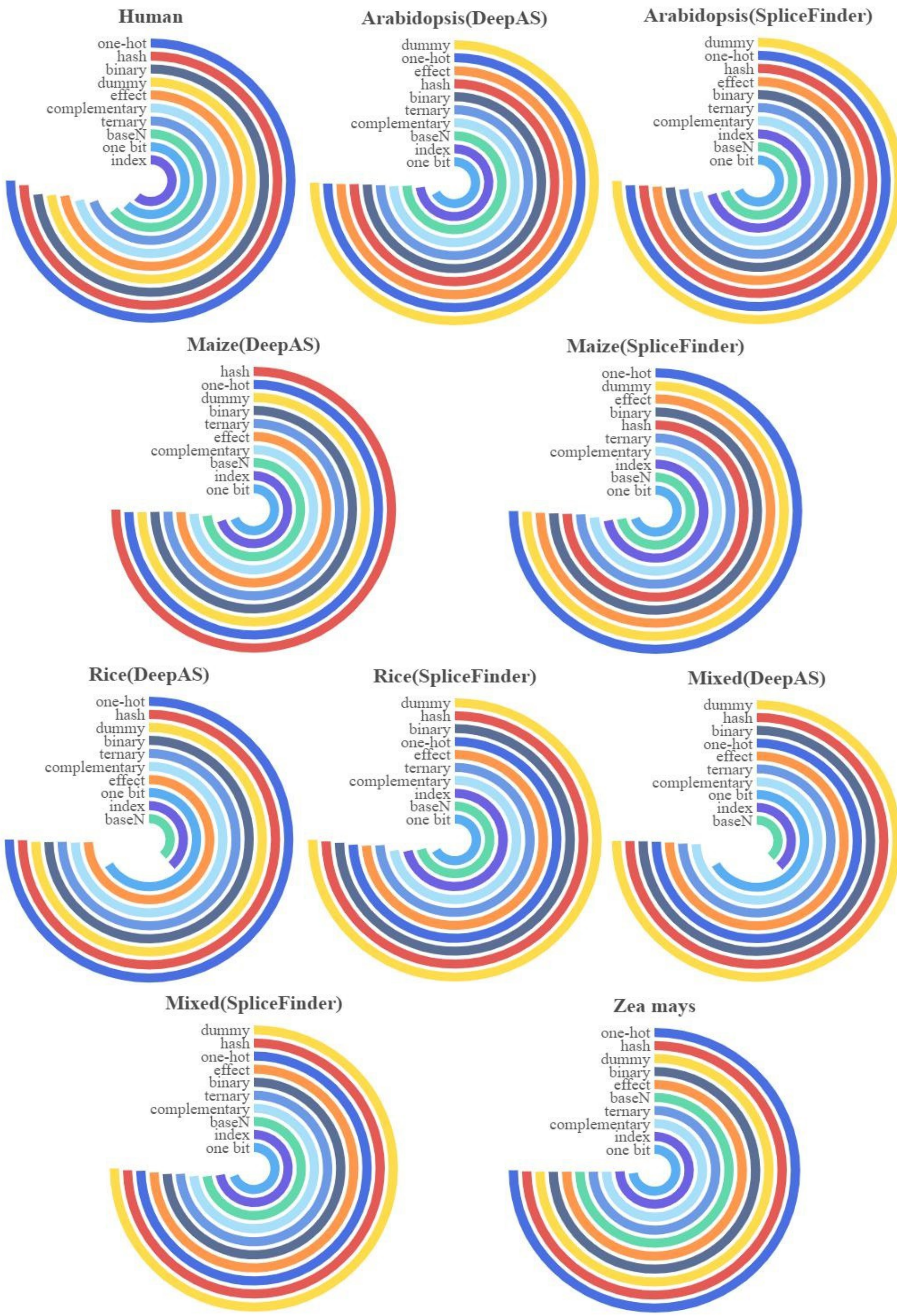


Fig. 3. Training process of models. In figure legend, At the top is the method to maximize the prediction accuracy (or PCC) of the model. A. Training process of human circRNAs using ten encoding methods in DeepCircRNA. B. The training process of *Arabidopsis*, maize, rice and their mixed datasets using ten encoding methods in DeepAS. C. The training process of *Arabidopsis*, maize, rice and their mixed datasets using ten encoding methods in SpliceFinder. D. Training process of *Zea mays* lncRNAs using ten encoding methods in PlncRNA-HDeep. E. Training process of *Saccharomyces cerevisiae* promoter using ten encoding methods in Evolution-gpu.

A



Human
one-hot
hash
binary
dummy
effect
complementary
ternary
baseN
one bit
index
Arabidopsis(DeepAS)
dummy
one-hot
effect
hash
binary
ternary
complementary
baseN
index
one bit
Arabidopsis(SpliceFinder)
dummy
one-hot
hash
effect
binary
ternary
complementary
index
baseN
one bit
Maize(DeepAS)
hash
one-hot
dummy
binary
ternary
effect
complementary
baseN
index
one bit
Maize(SpliceFinder)
one-hot
dummy
effect
binary
hash
ternary
complementary
index
baseN
one bit
Rice(DeepAS)
one-hot
hash
dummy
binary
ternary
complementary
effect
one bit
index
baseN
Rice(SpliceFinder)
dummy
hash
binary
one-hot
effect
ternary
complementary
index
baseN
one bit
Mixed(DeepAS)
dummy
hash
binary
one-hot
effect
ternary
complementary
one bit
index
baseN
Mixed(SpliceFinder)
dummy
hash
one-hot
effect
binary
ternary
complementary
baseN
index
one bit
Zea mays
one-hot
hash
dummy
binary
effect
baseN
ternary
complementary
index
one bit

B

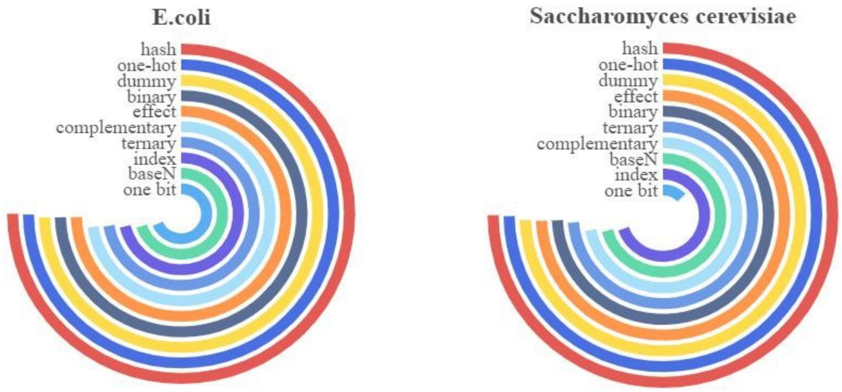
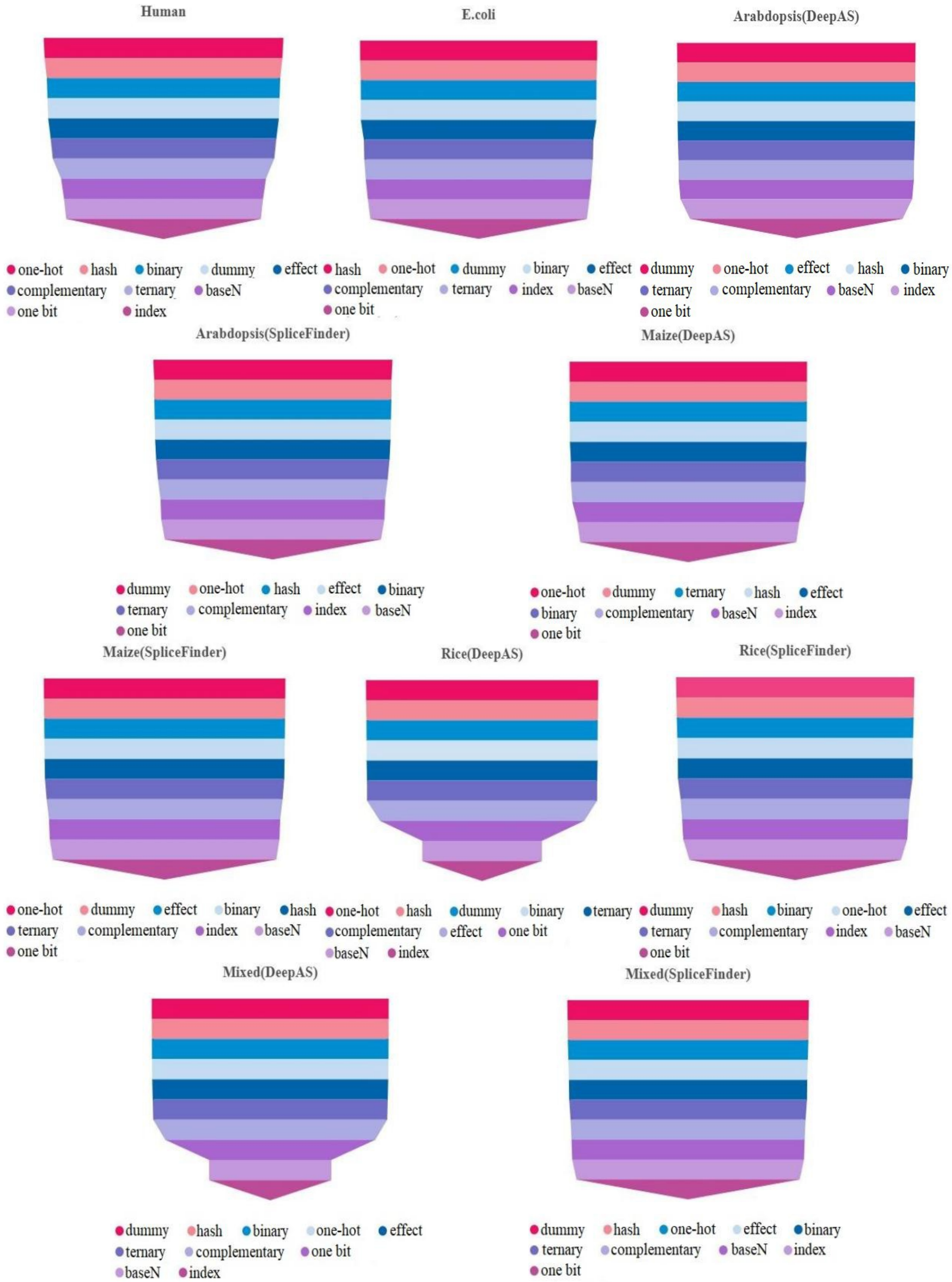


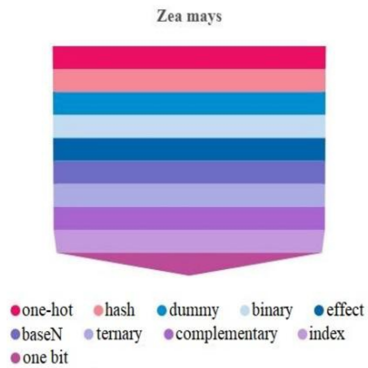
Fig. 4. Performance of ten enoding methods in different types of data. In figure legend, the method at the top has the highest prediction accuracy (or PCC). A. Performance of ten encoding methods in RNA datasets. Radial bar charts show that the prediction accuracy of dummy encoding and one-hot encoding is the highest. B. Performance of ten encoding methods in promoter datasets. Radial bar charts indicate that hash encoding can achieve the highest accuracy (or PCC).

A

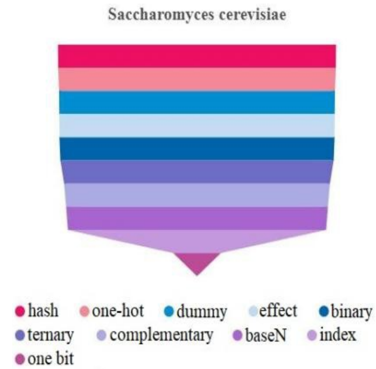


Human
one-hot hash binary dummy effect complementary ternary baseN one bit index
E.coli
hash one-hot dummy binary effect complementary ternary index baseN one bit
Arabdopsis(DeepAS)
dummy one-hot effect hash binary ternary complementary baseN index one bit
Arabdopsis(SpliceFinder)
dummy one-hot hash effect binary ternary complementary index baseN one bit
Maize(DeepAS)
one-hot dummy ternary hash effect binary complementary baseN index one bit
Maize(SpliceFinder)
one-hot dummy effect binary hash ternary complementary index baseN one bit
Rice(DeepAS)
one-hot hash dummy binary ternary complementary effect one bit baseN index
Rice(SpliceFinder)
dummy hash binary one-hot effect ternary complementary index baseN one bit
Mixed(DeepAS)
dummy hash binary one-hot effect ternary complementary one bit baseN index
Mixed(SpliceFinder)
dummy hash one-hot effect binary ternary complementary baseN index one bit

B



C



Supplementary Fig. 1. Performance of ten encoding methods in full sequence data and part sequence data. In this figure, the method with the highest prediction accuracy(or PCC) is at the top. A. Performance of all encoding methods in part sequence data. Funnel plots show that the best performance is dummy encoding, hash encoding and one-hot encoding. B. Performance of all encoding methods in full sequence data. Funnel plots show that one-hot encoding performs best, and hash encoding, dummy encoding follow closely. C. Performance of all encoding methods in data with both part and full sequence. Funnel plots show that hash encoding performs best, followed by one-hot encoding and dummy encoding.

Tables

Table 1. Datasets and corresponding models.

ID	Data	Data size	Data type	Model	Website
1)	human	65,966	circRNAs	DeepCircRNA	http://deepbiology.cn/DeepCircRNA
2)	*Arabidopsis*	20,000	linear RNAs	DeepAS, SpliceFinder	http://deepbiology.cn/DeepAS https://gitlab.deepomics.org/wangruohan/SpliceFinder
	Maize	20,000			
	Rice	20,000			
	Mixed	60,000			
3)	*Zea mays*	36,000	lncRNAs	PlncRNA-HDeep	https://github.com/Shujaatmalik/pcPromoter-CNN
4)	*E. coli*	5,720	promoter	pcPromoter-CNN	https://github.com/kangzhai/PlncRNA-HDeep
5)	*Saccharomyces cerevisiae*	100,000	promoter	Evolution-gpu	https://github.com/1edv/evolution/tree/master/manuscript_code/model/gpu_only_model

Table 2. Prediction accuracy using ten encoding methods in DeepCircRNA.

Encoding methods	Human
BaseN	0.8236
Binary	0.9378
Complementary	0.9070
Dummy	0.9327
Effect	0.9270
Hash	0.9557
Index	0.7825
One bit	0.7997
One-hot	**0.9630**
Ternary	0.8902

Table 3. Prediction accuracy using ten encoding methods in DeepAS.

Encoding methods	*Arabidopsis*	Maize	Rice	Mixed
BaseN	0.9445	0.9410	0.4997	0.4999
Binary	0.9535	0.9530	0.9655	0.9688
Complementary	0.9460	0.9530	0.9620	0.9578
Dummy	**0.9580**	0.9630	0.9665	**0.9696**
Effect	0.9555	0.9605	0.9615	0.9658
Hash	0.9550	0.9615	0.9695	0.9694
Index	0.9300	0.8994	0.4997	0.4999
One bit	0.8539	0.8774	0.8544	0.8592
One-hot	0.9565	**0.9635**	**0.9705**	0.9678
Ternary	0.9510	0.9625	0.9650	0.9618

Supplementary Table 1. Prediction accuracy using ten encoding methods in SpliceFinder.

Encoding methods	*Arabidopsis*	Maize	Rice	Mixed
BaseN	0.8795	0.9045	0.8980	0.9357
Binary	0.9273	0.9463	0.9498	0.9563
Complementary	0.9053	0.9293	0.9190	0.9434
Dummy	**0.9423**	0.9500	**0.9553**	**0.9695**
Effect	0.9310	0.9478	0.9415	0.9613
Hash	0.9343	0.9445	0.9528	0.9666
Index	0.8855	0.9103	0.9068	0.9304
One bit	0.8520	0.8795	0.8485	0.8960
One-hot	0.9355	**0.9510**	0.9453	0.9653
Ternary	0.9198	0.9420	0.9403	0.9542

Supplementary Table 2. Prediction performance of PlncRNA-HDeep using ten enoding methods.

Encoding methods	Zea mays			
	Accuracy	F1-score	Precision	Recall
BaseN	0.9578	0.9575	0.9498	0.9653
Binary	0.9590	0.9589	0.9477	0.9703
Complementary	0.9560	0.9567	0.9476	0.9639
Dummy	0.9596	0.9594	**0.9504**	0.9684
Effect	0.9586	0.9583	0.9503	0.9664
Hash	0.9603	0.9603	0.9463	0.9745
Index	0.9556	0.9553	0.9468	0.9639
One bit	0.9344	0.9328	0.9419	0.9238
One-hot	**0.9621**	**0.9620**	0.9497	**0.9746**
Ternary	0.9571	0.9571	0.9418	0.9729

Supplementary Table 3. Prediction performance of pcPromoter-CNN using ten enoding methods.

Encoding methods	Accuracy	AUC
BaseN	0.8075	0.9028
Binary	0.8587	0.9340
Complementary	0.8365	0.9151
Dummy	0.8593	0.9365
Effect	0.8586	0.9352
Hash	**0.8643**	**0.9368**
Index	0.8226	0.9086
One bit	0.7890	0.8659
One-hot	0.8633	0.9364
Ternary	0.8332	0.9237

Supplementary Table 4. PCC of Evolution-gpu using ten encoding methods.

Encoding methods	*Saccharomyces cerevisiae*
BaseN	0.8113
Binary	0.8526
Complementary	0.8237
Dummy	0.8548
Effect	0.8532
Hash	**0.8615**
Index	0.8001
One bit	0.1474
One-hot	0.8604
Ternary	0.8467

Supplementary Table 5. Accuracy of ten encoding methods in sisRNAs prediction model.

Encoding methods	*Arabidopsis*	Rice
BaseN	0.926	0.891
Binary	0.942	0.891
Complementary	0.945	0.814
Dummy	**0.947**	0.873
Effect	0.941	0.873
Hash	0.935	0.885
Index	0.939	0.865
One bit	0.946	0.635
One-hot	0.935	**0.896**
Ternary	0.939	0.880